Is extreme response style domain specific? Findings from two studies in four countries

Cabooter, Elke ; Weijters, Bert ; De Beuckelaer, Alain ; Davidov, Eldad

Abstract: Extreme response style (ERS) may bias responses and hamper the validity of conclusions in substantive research. ERS can be controlled for by using an additional (random) sample of response style indicators (i.e., a separate, random sample of survey items). There are two options to draw response style indicators to control for ERS: from only one versus from multiple domains. In two studies (four samples in total), this paper examines the domain dependency of ERS across three domains: consumer behavior, interpersonal relationships and politics. We find in the four samples repeated evidence suggesting that ERS has a domain specific component. This finding calls into question the (often encountered) assumption that it does not matter from which domains ERS measures are drawn.

DOI: https://doi.org/10.1007/s11135-016-0411-5

IS EXTREME RESPONSE STYLE DOMAIN SPECIFIC?

Findings from two studies in four countries

Elke Cabooter (corresponding author)
IÉSEG School of Management, Lille, France

Bert Weijters
Department of Personnel Management, Work and Organizational Psychology, Ghent University, Ghent, Belgium

Alain De Beuckelaer
Institute for Management Research, Radboud University, Thomas van Aquinostraat 1, P.O. Box 9108, 6500 HK Nijmegen, the Netherlands
School of Sociology and Population Studies, Renmin University of China, Beijing, P.R. China

Eldad Davidov
Institute of Sociology, and URPP 'Social Networks', University of Zürich, Switzerland
Institute of Sociology and Social Psychology, University of Cologne, Cologne, Germany

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in ***Quality & Quantity*** following peer review. It was first published online in this journal on 14 October 2016.

The definitive publisher-authenticated version is available online at Springer via

*http://dx.doi.org/*10.1007/s11135-016-0411-5

1. Introduction

According to classical test theory (Traub 1994), valid measurement implies the absence of systematic bias in respondents' answers. One major source of systematic bias in inter-individual responses to survey questions is extreme response style (Cheung and Rensvold 2000), abbreviated as ERS. ERS is the systematic tendency to preferentially use the extreme response categories for a range of survey items (Paulhus 1991). Its potential presence may be unfortunate for researchers who seek both valid and reliable measurements to reflect their theoretical constructs and test theories. Indeed, studies have amply demonstrated that comparisons of rating scale data (e.g., survey item scores on a 5-point agree/disagree scale) are particularly prone to *differences* in ERS across the groups to be compared (Baumgartner and Steenkamp 2001; Van Herk et al. 2004; Weathers et al. 2005; Weijters et al. 2010). The differences in ERS between members of different groups to be compared are directly responsible for a bias, now referred to as ERS bias. In the case of comparisons between individuals from different countries the severity of ERS bias is dependent on the particular set of countries involved in the comparison(s).

Several authors (Cheung and Rensvold 2000; De Jong et al. 2008) have demonstrated that ERS bias in particular threatens the validity and reliability of comparisons both within and across countries. It is very well known that even small deviations of a few percentage points in ERS may invalidate statistical data comparisons across countries (De Jong et al. 2008; Dolnicar and Grün 2007). Therefore, it is highly important to eliminate or control for ERS bias.

Luckily, different methodological procedures have been set up for diagnosing, controlling for, or correcting ERS bias such as the count procedure, an item response theory (IRT)-based method, latent class confirmatory factor analysis (LCFA), the representative indicators of response styles (RIRS) method, and the representative indicators response styles means and covariance structure (RIRMACS) method (see Van Vaerenbergh and Thomas 2013, for a detailed description of these methods). Van Vaerenbergh and Thomas (2013) recommend relying on a combination of these methods to adequately account for ERS bias. Except for LCFA and the IRT-based method, which are rather difficult to implement and require specialized software, the methods mentioned above rely on an "easy implementation", that is, the use of a *separate* set of survey items randomly drawn from multiple domains to quantify each respondent's proneness to ERS. This approach ensures that the ERS indicators capture style and not content (Baumgartner and Steenkamp 2001; De Beuckelaer et al. 2010; Kieruj and Moors 2013; Weijters et al. 2008). The recommendation to sample survey items from various domains relies on the assumption that ERS is not domain specific. In the current research, which includes two studies, we test this assumption. Specifically, in study 1, we administered a survey and replicated the survey administration so that we obtained undergraduate student samples studying in programs conducted in English in three countries (Belgium, the Netherlands, and Germany). In a second study, we composed a non-student sample (USA) and administered the same survey. On the basis of these two studies we were able to measure ERS across three content domains.

In the following, we will first discuss potential problems in the measurement of ERS bias. Next, we will present our two studies, the data used, the measurement, and

provide empirical findings which suggest that ERS is partly domain specific. Then we will conclude with some final remarks and possible implications of the findings.

1.1 Measuring ERS and potential obstacles

To measure ERS, typically a separate set of survey items is needed in addition to the survey items that are already included in the study (which we refer to as ERS indicators **(Editorial instruction: Footnote 1 BEFORE the end of the sentence)**. Every single ERS indicator informs the researcher as to whether or not the respondents under study have given the focal response, in our case, an extreme answer to the survey item. If an extreme response is given by a respondent, the ERS indicator takes on the value of one. Otherwise, the ERS indicator takes on the value of zero. Obviously, multiple ERS indicators are needed to quantify each respondent's proneness to ERS. In a typical case in which the researcher is not interested in the response to a specific ERS indicator but rather in the general response pattern, the information contained in the complete set of ERS indicators needs to be summarized in an adequate individual-level summary statistic. An adequate summary statistic is expressed as the proportion of ERS indicators (in the set) evoking extreme responses. These summary statistics then provide an estimate of the extent to which an individual respondent is prone to ERS.

As mentioned above, various researchers make use of ERS indicators which are randomly sampled from multiple domains, that is, they rely on a heterogeneous set of substantive domains (see, e.g., Weijters et al. 2008; Weijters et al. 2010). A key question in this regard is how one should interpret the notion of "multiple domains". If ERS bias is consistent across different domains, then sampling heterogeneous survey items from

different substantive domains would be fine. However, Hui and Triandis (1989) found that respondents who are more involved with the subject of the survey tend to have more outspoken opinions and tend to select extreme response categories more frequently than respondents who are less involved. Van Dijk et al. (2009) found that measures of ERS as obtained from different domains display low associations with each other. Indeed, various domains may invoke different levels of emotional involvement or social desirability and, consequently, different response bias tendencies (Hirschfeld and Gelman 1994; Hui and Triandis 1989; Steenkamp et al. 2010). Therefore, we expect that randomly selected indicators from *the same* domain are likely to invoke *the same* or a similar pattern of response bias, including ERS response bias. If it turns out that ERS is domain dependent, it is clear that one should refrain from sampling ERS indicators from a broad universe of substantive domains. Instead, it would be clearly preferable to sample ERS indicators from a more narrowly defined, relevant universe.

Empirical evidence so far also seems to indicate that summary statistics of ERS may vary depending on the *culture* under study from which the ERS indicators are derived. Thus, the literature emphasized that due to differential levels of ERS across cultures, cross-cultural data may not be comparable. Different levels of ERS may be a major source of this problem (see, e.g., Davidov et al. 2014). Therefore, comparing data over cultures is not warranted unless differences in the use of ERS (in a given culture or country) are diagnosed and adequately controlled for (Steenkamp and Baumgartner 1998).

Given the considerations above, we formulate the following two hypotheses:

(H1) We expect measures of ERS to be domain specific; (H2) We expect to find support for this expectation in various country-specific samples, including one non-student sample. Hypothesis 2 is added in order to include a 'robustness check' for our empirical findings regarding Hypothesis 1. For study 1, we will conduct the test in similar (undergraduate student) samples drawn from different countries. To avoid dealing with differential (country-specific) levels of ERS due to different levels of English command we relied exclusively on students enrolled in an English-taught program. In study 2, we will conduct the test, once more, for a non-student sample in another country (the USA). As such, we will try to generalize the findings obtained from Study 1.

2. Study 1

2.1. Method

2.1.1. Participants

This study relies on a convenience sample (N = 349) of undergraduate students in three (rather similar) West European countries: Belgium (Flanders only, n = 134; 67.2% women; M_{age} = 21.91, SD_{age} = 2.78), the Netherlands (n = 120; 55.8% women; M_{age} = 23.92, SD_{age} = 3.43), and Germany (n = 95; 74.7% women; M_{age} = 25.09, SD_{age} = 3.87).

All students included in these samples were country nationals with at least one parent having the same nationality. They were all enrolled in English-taught educational programs, and they all passed the international "Test of English as a Foreign Language" (TOEFL). Thus, we expect that all students have a good command of the English language. Completing a survey in a language which is not one's own native tongue (this is the case in all three samples) but in a language that one does master is important,

because completing surveys in one's own language increases the amount of extreme responding (Harzing 2006). As such, the decision to make use of a common language, English, is instrumental as indicated earlier in establishing data comparability across the three countries (Davidov and De Beuckelaer 2010; Weijters et al. 2013). All students in the three separate country samples indicated (in their response to a separate question) that they did not experience any considerable difficulty in completing the survey.

2.1.2. Survey content

In addition to basic demographic data including age and gender, the survey included a set of ERS indicators sampled from various domains. The set of ERS indicators contained three sets of 16 (randomly sampled) survey items which served as the ERS indicators. Each set represented one of the following three domains: (1) consumer behavior (CB); (2) interpersonal relationships (IR); and (3) politics (POL). The reason for selecting these domains is twofold: (1) Each domain represents an important research area in the academic literature (with 1.71, 2.20, and 2.17 million hits on Google Scholar, respectively, signifying that a large number of researchers have devoted much time and effort on these subjects), and (2) several scale handbooks (see below) are available for each domain. These handbooks offer the appropriate sampling frames to select survey items. The selection of survey items is based on the principle of random sampling. Once selected for inclusion in the final survey (after approval, see explanation below), the fixed set of randomly sampled survey items is offered to all respondents in the same fixed order.

Survey items reflecting consumer behavior have been randomly chosen from the *Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research* (Bearden and Netemeyer 1999) and *Marketing Scales Handbook: A Compilation of Multi-Item Measures* (Bruner et al. 2005). Survey items representing interpersonal relationships (i.e., love, friendship, and family relations) have been randomly sampled from various sources, namely, the *Measurement of Love and Intimate Relations* (Tzeng 1993), *A Bare Bones Guide to the Acquaintance Description Form* (Wright 1997), *Handbook of Sexuality-Related Measures* (Davis 1998), and *Handbook of Family Measurement Techniques* (Perlmutter et al. 2001). Finally, survey items on political matters (including attitudes) have been randomly selected from the *Measures of Political Attitudes* (Robinson et al. 1969). All survey items were scored on a fully labeled five-point disagree/agree scale with the following labels: *strongly disagree*, *disagree*, *neither disagree nor agree*, *agree*, *strongly agree*.

Respondents' comprehension and usability of the sampled survey items were determined on the basis of two analogous consecutive pretests using a sample of 20 students. In the first pretest, 20 respondents from the Netherlands were asked to complete the initial survey including 48 (i.e., 16 x 3) randomly drawn survey items and to indicate all words/terms that were unclear to them. Based on the results from this first pretest, seven survey items were slightly adapted to facilitate respondents' comprehension (e.g., replacing the unfamiliar word "genuine" with the more familiar word "real"). In addition, 21 survey items were discarded. The major reasons for removing the 21 survey items were: (a) the fact that students were not really able to provide a score for the survey item (e.g., they indicated having no opinion on the effectiveness of the United Nations), or (b)

the data collected showed that the survey item reflected a general opinion (i.e., with scores showing near-to-zero variation) rather than one's personal opinion. Because the final survey needed to include 16 randomly sampled survey items from each domain (i.e., 48 survey items in total), a second procedure was required to identify 21 new survey items to replace those that had been removed after the first pretest. This second pretest procedure consisted of two steps: (1) 36 new survey items were randomly drawn from the three relevant domain (i.e., 12 from each domain), and (2) a new sample of 20 students from the Netherlands completed these 36 survey items and indicated all words/terms that were unclear. Based on their responses, 21 new survey items (from the survey item pool of 36) have been selected as replacements. After this step, the final survey comprised 48 randomly sampled survey items which included 16 survey items for each of the three domains. All 48 survey items are listed in the Appendix. In support of the intended content heterogeneity, the averaged correlation between survey items was $r = 0.029$ for consumer behavior, $r = -0.003$ for interpersonal relationships, and $r = 0.026$ for political matters, for a summary see Figure 1.

2.1.3 Summary statistics of ERS

We pursued the following approach to address the question of whether or not, and to what extent ERS is domain specific. As described above, the data contained responses to 48 survey items, also referred to as ERS indicators, from three domains (with 16 survey items per domain). An ERS score was computed by giving each extreme survey item response a score of one (instead of the default zero score). To indicate a respondent's proneness to ERS for a particular domain of survey items, the information contained in

all 16 ERS indicators reflecting the domain under study was summarized in two summary statistics per domain. Each domain specific summary statistic was based on a different set of 8 out of the 16 ERS indicators linked to that domain **(Editorial instruction: Footnote 2 BEFORE OR AFTER the end of the sentence)**. Obviously, the two summary statistics together (jointly) summarize all ERS indicators representing that domain. Across all three domains, six ERS summary statistics were computed (two ERS summary statistics in the CB domain, two ERS summary statistics in the POL domain, and two ERS summary statistics in the IR domain). We specifically split the 16 survey items in each domain by assigning the odd numbered survey items to the set of items used to derive an "odd" ERS summary statistic and all even numbered survey items were assigned to the set of items to derive an "even" ERS summary statistic. Next, we verified the generalizability of our findings by studying the effect per domain by alternatively using two random subsets of eight survey items.

Having two rather than one ERS summary statistic per domain allows us to assess whether ERS summary statistics from the same domain correlate more strongly with one another than with ERS summary statistics from other domains. Thus, the individual variance of the ERS summary statistics indicates the presence of random measurement variance; the shared variance of the two ERS summary statistics (within each domain) reflects domain specific ERS whereas the shared variance of the six summary statistics (across domains) reflects non domain specific ERS. As such, we can control for measurement error. This comparison between summary statistics can be conceived of as a multitrait-multimethod (MTMM) procedure (Eid et al. 2006), where the three domains

can be considered as reflecting distinctive traits of interest, and the two ERS summary statistics (based on odd vs. even survey items) as two distinctive methods.

2.2. Results

Table 1 presents the correlation matrix for the overall sample and for each country separately, displayed as an MTMM matrix (Eid et al. 2006). The MTMM matrix reveals three important findings. First, all correlations are significantly higher than zero ($p < .05$). The fact that all ERS summary statistics measured for different domains are positively related suggests the presence of a general ERS variance. Second, the monotrait-heteromethod correlations (i.e., within-domain correlations between the ERS summary statistics based on odd vs. even survey items) are consistently higher than the heterotrait correlations (i.e., correlations between ERS summary statistics representing different domains), and this observation holds for the overall sample as well as for each of the three countries. The repeated character of this finding suggests that part of the ERS variance is domain specific. Since the survey items in each domain were sampled from many different scales, construct specific content cannot account for this domain specific ERS variance. Third, measures of internal consistency are highest for the POL domain. Apparently, ERS indicators (i.e., survey items) in this content domain contain more shared ERS variance than is the case with CB and/or IR.

As a more formal test of the covariance structure of the ERS summary statistics, the structural equation model (SEM) presented in Figure 2 is fitted to the data in Mplus 7.4 using the default ML estimator. ERS is modeled as a latent variable on which all six ERS summary statistics load (i.e., two summary statistics for each domain). This latent

variable (or “factor”) represents a generalized ERS factor, that is, ERS across domains. Two summary statistics from the same domain have equal loadings (since they are based on two interchangeable sets of survey items from the same domain). In addition, two summary statistics from the same domain have correlated residual terms. The idea is that, if ERS has a domain specific component, summary statistics representing the same domain will show a significant residual correlation (after controlling for the general ERS). All ERS manifest variables loaded considerably on the general ERS latent variable, reflecting the general common variance across domains. Model fit indices for this model are reported in Table 2 (in the row labeled 1a. base model).

To further investigate the extent to which ERS is or is not equivalent across domains, we test additional parameter constraints. Model 2 imposes equal factor loadings across the three domains (i.e., the factor loadings for CB, IR and POL are equal). Model 3 imposes equal residual covariance terms across the three domains (i.e., the residual covariance terms for CB, IR and POL are equal). Model fit indices for the resulting models are reported in Table 2. Chi^2 difference tests (see the last three columns of Table 2) show that imposing cross-domain factor loading equality results in a significant deterioration of model fit and should be rejected. Imposing cross-domain residual covariance equality does not lead to a significant deterioration in fit and cannot be rejected. Parameter estimates for the final model (i.e., Model 3), are reported in Table 3.

Cross-group invariance tests demonstrate that the factor loadings and the residual covariance terms are invariant *across samples* (see Model 1b and model 1c in Table 2) implying the similarity of the three samples we used; we therefore constrain these parameters to equality across the three groups. Further model testing shows that the factor

loadings are significantly different across domains (see Model 2), but the residual covariance term is not significantly different across domains.

Closer inspection of the parameter estimates and the model tests indicate the following. First and most importantly, for all countries under study and for all three domains, there is a domain specific ERS component, in that the residual covariance term between two summary statistics from the same domain is statistically significant (with standardized estimates ranging from .23 to .28; see Table 3). These significant residual correlations indicate the presence of domain specific ERS variance because, after controlling for general ERS (i.e., the general ERS factor), the two summary statistics from the same domain share (additional) ERS variance whereas two summary statistics from different domains do not share (additional) ERS variance.

Second, this residual covariance term is not significantly different across domains. This means that two summary statistics based on two sets of survey items from the same domain tend to show similar levels of relatedness after correcting for a general (non-domain specific) ERS factor. However, third, the factor loadings of summary statistics on the general ERS factor are sometimes different *across domains* within samples. Specifically, the POL summary statistics have significantly stronger loadings on the general ERS factor than do the CB summary statistics. We additionally find that the CB loadings are lower than the IR loadings, and that the IR loadings are lower than the POL loadings. These differences in loadings across domains suggest that POL survey items are more prone to a general ERS than are CB survey items, with IR survey items situated in between. These results indicate that both Hypothesis 1 (domain specificity of response

styles) and Hypothesis 2 (Hypothesis 1 is supported in various samples) receive empirical support.

The calculation of ERS summary statistics was based on a so-called “odd-even” split (i.e., the two summary statistics representing each domain were calculated using odd and even ERS indicators [survey items], respectively). An odd-even split is commonly used to assess split-half reliability, but many other random splits are possible. To explore the extent to which the above results are specific to the odd-even split, from the existing data, 400 datasets were generated in which the 16 ERS indicators (survey items) per domain were split in different random ways **(Editorial instruction: Footnote 3 BEFORE the closing dot)**. Using the MONTECARLO analysis module in Mplus, the same models as presented above were fitted to the data at hand and the averaged parameters and standard errors were reported.

For the MTMM model (see Figure 2), model fit is generally good across the 400 datasets (see Table 2.) The averaged results for the parameter estimates are shown in Table 3. The results demonstrate that the domain specificity of ERS is not dependent on the way the ERS summary statistics are combined. In sum, this additional simulation research provided even stronger support for Hypothesis 1 and Hypothesis 2.

3. Study 2

In Study 1, the purpose of the study was to examine whether ERS is domain specific. In Study 2, a single country study, we aim to replicate the findings of Study 1 for respondents from the USA. We have chosen for a USA sample as all citizens have English as a native language and since many studies in social sciences use USA data.

Study 2 also helps us to assess whether our earlier study findings generalize to a non-student sample.

3.1.Method

3.1.1. Participants

One hundred forty nine participants recruited from the Amazon Mechanical Turk (MTurk) database (73 men, 76 women; M_{age} = 36.4 years, SD = 11.9; American nationality = 89.9%; English as native language = 100%) took part in this study and were paid $0.40 each for participation.

3.1.2. Survey content

Similar as in study 1, the questionnaire contained three sets of 16 (randomly sampled) survey items. Each set represented one of the following three domains: (1) consumer behavior (CB); (2) interpersonal relationships (IR); and (3) politics (POL). In addition, some demographic information was collected including age, gender and nationality.

In support of the intended content heterogeneity, the averaged correlation between survey items was r = 0.064 for consumer behavior, r = 0.021 for interpersonal relationships, and r = 0.061 for political matters, see Figure 3.

3.2. Results

Table 4 presents the correlation matrix for the overall sample and for each country separately, displayed as an MTMM matrix (Eid et al. 2006). As in study 1, the correlations indicate that ERS has a domain specific component. More specifically, the

within domain correlations are higher than the between domain correlations. Similar as in study 1, the internal consistency is highest for POL. The common variance implies that, there is substantial correlation between ERS items also across domains.

To further investigate the extent to which ERS is or is not equivalent across domains, we test additional parameter constraints similarly to Study 1. Model 2 imposes equal factor loadings across the three domains (i.e., the factor loadings for CB, IR and POL are equal). Model 3 imposes equal residual covariance terms across the three domains (i.e., the residual covariance terms for CB, IR and POL are equal). Model fit indices for the resulting models are reported in Table 2. Chi-square difference tests (see last three columns in Table 2) indices show that imposing cross-domain factor loadings equality results in a significant deterioration of model fit and should be rejected. Imposing cross-domain residual covariance equality does not lead to a significant deterioration in fit and cannot be rejected. Parameter estimates for the final model (i.e., Model 3) are reported in Table 3.

In sum, we were able to replicate the findings of study 1 also in Study 2. First, there seems to be a domain specific ERS component. Second, in both studies the POL summary statistics have stronger loadings on the ERS factor than do the CB summary statistics. However, compared to Study 1 where we find that the CB loadings are lower than the IR loadings and the IR loadings are lower than the POL loadings, in Study 2 those differences are less outspoken (though the difference in loadings between CB and POL is significant as well). Overall, these differences in loadings across domains suggest that POL survey items are more prone to a general ERS than are CB survey items, with

IR items situated in between. In sum, we found that study 2 further supports both Hypothesis 1 and Hypothesis 2.

4 Conclusions and discussion

To control for ERS in surveys, contemporary methodological guidelines recommend the inclusion of a separate set of survey items sampled from multiple domains (e.g., Van Vaerenbergh and Thomas 2009). This set of survey items in these studies is often a representative set from a universe of ERS indicators. Other research findings (Hui and Triandis 1989, Van Dijk et al. 2009) have casted doubts about an important assumption made, namely the assumption that it does not matter from which domains ERS measures are drawn.

Our findings suggest that ERS bias varies depending on the *domain* from which the survey items are randomly sampled. In other words, different domains may induce different patterns of response bias in general, and ERS response bias in particular.

The results of the two studies we conducted may have implications for research practice. Sampling survey items from a broad universe of items may not necessarily generate an optimal set of ERS indicators, especially if the substantive questions included in the survey are taken from one domain (or just a few domains). Such situations are typically encountered when using or analyzing data from, for example, social scientific theme-related surveys such as the annual International Social Survey Program (ISSP). A good alignment of the boundaries of the universe of (potential) ERS indicators to be sampled from (i.e., one's survey item bank) and the substantive content of the survey may be required when the goal is to adequately control for respondents' ERS as manifested in

the corresponding substantive domain in the survey. Both our studies show that ERS bias is not identical across different domains. In other words, both studies provide evidence that extreme responding may be at least partly domain specific. Consequently, we advise researchers to consider including a separate set of 15 survey items of the domain under subject to control for ERS influences (Weijters et al. 2009). In case a survey focuses on multiple domains one may consider including in the survey, for each domain, a separate set of 15 survey items to serve as ERS indicators. However, we acknowledge that this approach is resource consuming, especially when the number of domains covered in the survey is large. Yet, such efforts are achievable for a survey in one domain.

Just as any other empirical research, our research has a number of weaknesses. These weaknesses do, however, offer opportunities for further research. First, the current research used student samples from three European countries in the first study. Even though study results were replicated for a non-student USA sample in our second study, additional research could examine the domain effect in other countries with larger non-student samples, because response tendencies to scale formats are likely to vary across nations or cultures (Diamantopoulus, Reynolds, & Simintiras 2006). Thus, further research could replicate the study on other (same population) samples or for other countries (i.e., other populations).

In addition, we could have chosen to administer the survey in our Study 1 in one's own native language. However, in this research we have chosen to avoid translating the survey to one's own language given the large number of potential confounding factors such as translation mistakes, differences in language use and language competence (Harzing 2006), or potential subtle differences in the meaning of the scale items or labels

in different languages (Weijters et al. 2013; De Langhe et al. 2011; see also Podsakoff et al. 2003). The inclusion of students with a similar command of the English language (i.e., our research strategy) is likely to circumvent problems related to confounding. In addition, applying one's native language in a questionnaire may possibly lead to an increased level of extreme responding (Harzing 2006). Finally, Study 2, which was based on a non-student sample of study participants responding in their own native language (English), replicated the findings obtained from our student samples.

This research is also limited in that ERS indicators were only sampled from three (albeit clearly distinct and important) domains. Broadening the scope by adding more domains might be instrumental in identifying domain clusters that tend to produce very similar estimates of ERS. In addition, one might also search for ERS responding patterns that are specific for certain domain clusters. For instance, item wording characteristics can help explain the domain specificity of ERS. As can be seen in appendix, items in the POL domain do typically not contain pronouns in the first person singular, whereas items in IR do. Items in the POL domain also tend to be longer than those in other domains. Also other aspects can perhaps explain the domain specificity of ERS, namely Smith (2004) suggested that domains can differ in terms of personal relevance. As such, further research can try to group different domains according to item wording and/or their level of personal relevance and determine whether ERS summary statistics differ depending on these aspects.

In sum, this research has highlighted the importance of addressing a practical question, namely, from which domains should ERS indicators be sampled? Our survey, replicated in four samples, revealed that, to allow for a more optimal account of ERS, one

should consider also the possibility of including indicators measuring ERS from domain(s) matching the domain(s) under study as closely as possible, because ERS from different domains behaved differently. Further studies are needed to replicate our results on different domains and on more representative population samples covering additional nations and domains. Yet, our findings call into question the current procedures to address ERS using only items from diverse domains.

Footnotes

In terms of the terminology used, a cautionary note is warranted. In the context of their study, Weijters et al. (2008) use the term “response style indicator” to refer to an individual-level summary statistic denoting the number of times a focal response style occurs in a randomly drawn set of survey items (in their study: drawn from a broad variety of domains). However, as mentioned below, in our study every single survey item included in the random set (in this study: drawn from a more narrowly defined domain) is referred to as a response style indicator. As a consequence, in Weijters et al. (2008), a response style indicator represents a count variable (in our study a count variable would qualify as a “summary statistic”), whereas in our study a response style indicator can only take on a value of zero (i.e., response style is not present) or a value of one (i.e., response style is present), depending on what answer the respondent provided to the survey item under consideration.

[2] For studies involving an assessment of response styles, Weijters et al. (2008) proposed a criterion of using 10 to 14 response style indicators to quantify summary statistics for response styles. Closer inspection of the empirical results from a sensitivity analysis on which this criterion has been proposed, clearly shows that this criterion applies primarily to response styles other than ERS (i.e., acquiescence, disacquiescence, and midpoint response style). The sensitivity analysis conducted by Weijters et al. (2008), showed that ERS is—by far—the most stable and reliable response style, and ERS measures tend to demonstrate strong construct reliability and reliable variance estimates with as few as five survey items used to compute an ERS summary statistic. Hence, our

current use of eight survey items per summary statistic is consistent with response style measurement guidelines.

[3] We used the random number generator in MS Excel to obtain 400 different ways of splitting the 16 ERS indicators into two halves forming the two ERS summary statistics; e.g., in one dataset, the first CB ERS summary statistic might consist of CB survey items 1,4,5,7,10,11,12,16 while the other ERS summary statistic consists of CB survey items 2,3,6,8,9,13,14,15).

Tables and figures

Table 1 MTMM correlation matrix of ERS summary statistics based on odd versus even survey items

			Odd survey items			Even survey items		
			CB, odd	IR, odd	POL, odd	CB, even	IR, even	POL, even
Overall sample	Odd	CB, odd	*.37*					
		IR, odd	.34	*.36*				
		POL, odd	.34	.41	*.63*			
	Even	CB, even	**.47**	.30	.42	*.45*		
		IR, even	.32	**.57**	.43	.35	*.53*	
		POL, even	.29	.44	**.62**	.42	.45	*.64*
the Netherlands	Odd	CB, odd	***.42***					
		IR, odd	.37	***.45***				
		POL, odd	.39	.45	***.58***			
	Even	CB, even	**.54**	.42	.51	***.39***		
		IR, even	.41	**.57**	.38	.47	***.53***	
		POL, even	.33	.48	**.73**	.42	.44	***.63***
Belgium	Odd	CB, odd	***.31***					
		IR, odd	.27	***.24***				
		POL, odd	.32	.36	***.64***			
	Even	CB, even	**.41**	.29	.38	***.42***		
		IR, even	.20	**.52**	.43	.34	***.56***	
		POL, even	.26	.42	**.61**	.38	.43	***.63***
Germany	Odd	CB, odd	***.36***					
		IR, odd	.30	***.25***				
		POL, odd	.33	.38	***.63***			
	Even	CB, even	**.43**	.10	.38	***.51***		
		IR, even	.32	**.59**	.42	.20	***.45***	
		POL, even	.22	.31	**.52**	.41	.40	***.63***

Note. CB = consumer behavior, IR = interpersonal relationships, POL = political issues. All correlations are significantly higher than zero ($p < .05$). The numbers in gray italics are KR20 coefficients (based on the eight binary indicators that take on a value of one only if a survey item response is one or five, and a value of zero otherwise). The numbers in bold represent monotrait-heteromethod correlations. The remaining numbers are heterotrait correlations.

Table 2 Standardized parameter estimates for the correlated residuals model (cf. Figure 2)

			Model fit					Chi² difference test			
Study	Data	Model	Chi²	DF	RMSEA	CFI	TLI	SRMR	Ref.	Chi²	DF	p
Study 1 (3 countries)	O/E	1a. Base model	29.47	27	0.028	0.996	0.993	0.042				
		1b. Metric invariance	34.59	31	0.032	0.994	0.991	0.051	1a	5.12	4	0.275
		1c. Metric invariance and residual cov. Invariance	37.05	37	0.003	1.000	1.000	0.053	1b	2.46	6	0.873
		2. Equal loadings across domains	60.52	39	0.069	0.964	0.959	0.084	1c	23.4	2	< .001
		3. Equal residual covariance across domains	40.74	39	0.020	0.997	0.997	0.057	1c	3.69	2	0.158
	MI	1a. Base model	42.83	27	0.064	0.971	0.952	0.058				
		1b. Metric invariance	48.08	31	0.063	0.969	0.955	0.065	1a	5.25	4	0.263
		1c. Metric invariance and residual cov. Invariance	51.30	37	0.051	0.973	0.969	0.068	1b	3.22	6	0.781
		2. Equal loadings across domains	74.99	39	0.088	0.935	0.925	0.094	1c	23.69	2	< .001
		3. Equal residual covariance across domains	55.14	39	0.054	0.970	0.966	0.072	1c	3.85	2	0.146
Study 2 (US)	O/E	1. Base model	11.07	9	0.039	0.996	0.993	0.022				
		2. Equal loadings across domains	22.32	11	0.083	0.978	0.971	0.069	1	11.25	2	0.004
		3. Equal residual covariance across domains	14.78	11	0.048	0.993	0.990	0.027	1	3.71	2	0.156
	MI	1. Base model	12.90	9	0.045	0.99	0.98	0.03				

				0	7	8				
		1		0.97	0.96	0.07		11.7		0.00
2. Equal loadings across domains	24.69	1	0.089	2	2	6	1	9	2	3
3. Equal residual covariance across		1		0.98	0.98	0.04				0.11
domains	17.16	1	0.053	7	3	1	1	4.26	2	9

Note1. O/E = odd-even split; MI = multiple imputation; cov. = covariance term. For the MI results, reported fit indices are the mean across 400 datasets.
Note2. CB = consumer behavior, IR = interpersonal relationships, POL = political issues.

Table 3 Parameter estimates (SE) for the correlated residuals model (cf. Figure 2)

		Standardized loading			Residual correlation
		CB	IR	POL	
Study 1 (3 countries)	odd-even	0.494 (0.052) $_{a}$	0.600 (0.050) $_{b}$	0.701 (0.045)$_{c}$	0.277 (0.039)
	MC	0.484 (0.052) $_{a}$	0.575 (0.050) $_{b}$	0.716 (0.045)$_{c}$	0.229 (0.038)
Study 2 (US)	odd-even	0.747 (0.036) $_{a}$	0.773 (0.034) $_{a, b}$	0.794 (0.032) $_{b}$	0.242 (0.052)
	MC	0.742 (0.037) $_{a}$	0.745 (0.036) $_{a, b}$	0.800 (0.031) $_{b}$	0.178 (0.053)

Note1. All parameters are significant at $p < .01$. Parameter estimates with the same subscript are not significantly different from one another (at $p < .05$). In study 1, parameters were invariant across groups (see Table 2). For the MC results, reported parameters and standard errors are mean estimates across 400 datasets.

Note2. CB = consumer behavior, IR = interpersonal relationships, POL = political issues.

Table 4 MTMM correlation matrix of ERS summary statistics based on odd versus even survey items, USA sample

			Odd survey items			Even survey items		
			CB, odd	IR, odd	POL, odd	CB, even	IR, even	POL, even
USA sample	Odd	CB, odd	***.67***					
		IR, odd	.53	***.58***				
		POL, odd	.61	.63	***.66***			
	Even	CB, even	**.69**	.62	.60	***.63***		
		IR, even	.49	**.69**	.61	.53	***.59***	
		POL, even	.55	.63	**.69**	.62	.68	***.69***

Note. CB = consumer behavior, IR = interpersonal relationships, POL = political issues. All correlations are significantly higher than zero ($p < .05$). The numbers in gray italics are KR20 coefficients (based on the eight binary indicators that take on a value of one only if a survey item response is one or five, and a value of zero otherwise). The numbers in bold represent monotrait-heteromethod correlations. The remaining numbers are heterotrait correlations.

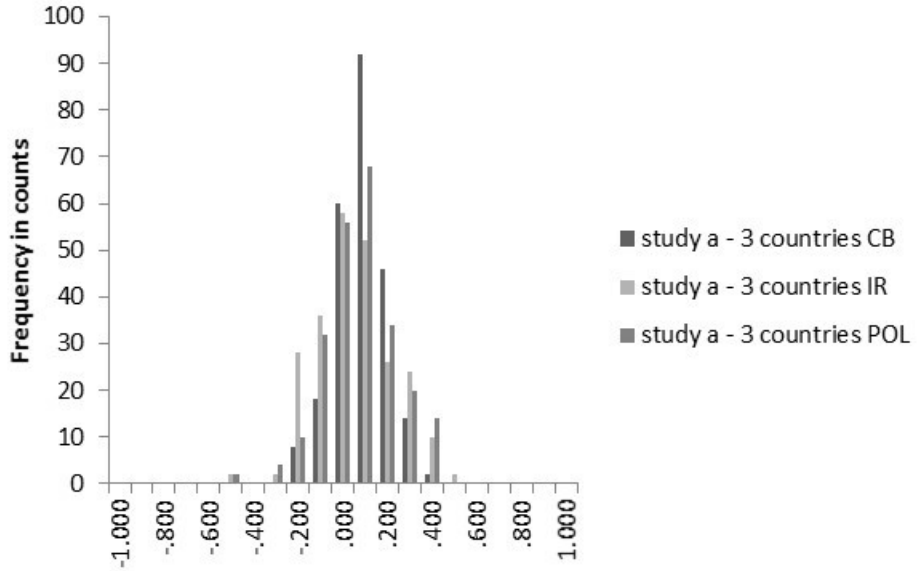

Figure 1 The frequency distributions of inter-item correlations by domain.

Note. CB = consumer behavior, POL = political issues, IR = interpersonal relationships.

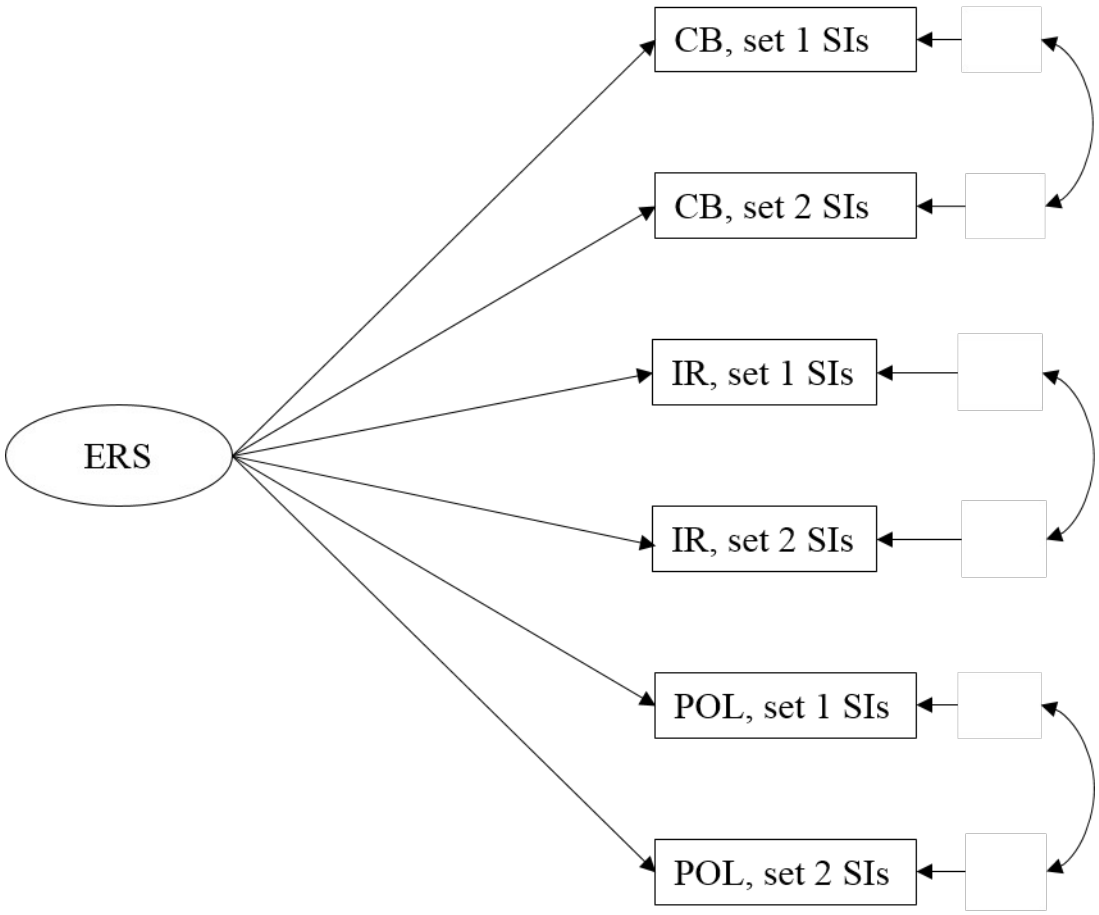

Figure 2 MTMM structural equation model.

Note. CB = consumer behavior, POL = political issues, IR = interpersonal relationships.

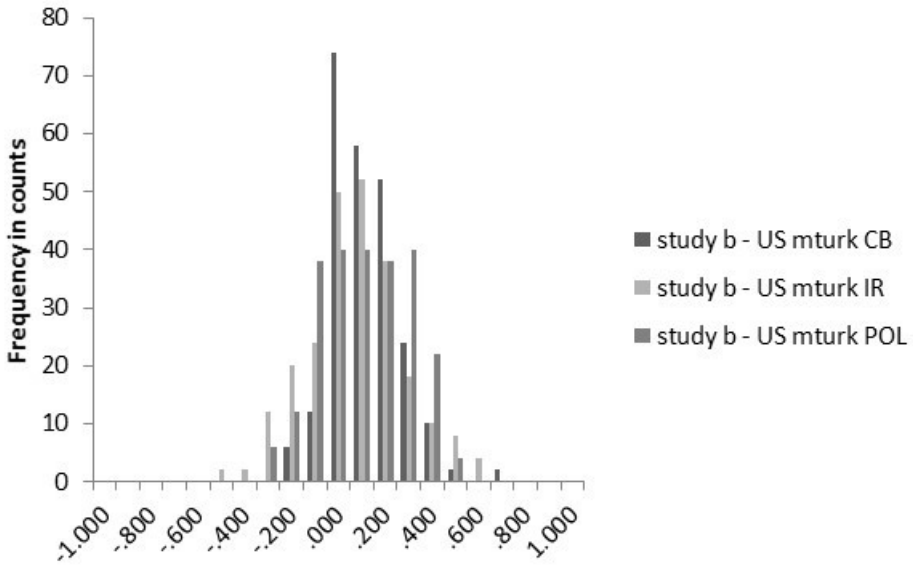

Figure 3 The frequency distributions of inter-item correlations by domain USA.

Note. CB = consumer behavior, POL = political issues, IR = interpersonal relationships.

References

Baumgartner, H., Steenkamp, J.-B. E. M.: Response styles in marketing research: a cross-national investigation. Journal of Marketing Research, **38**, 143-156 (2001) doi:10.1509/jmkr.38.2.143.18840.

Bearden, W. O., Netemeyer, R.G.: Handbook of marketing scales: multi-item measures for marketing and consumer behavior research. (2nd ed.). Sage, London, U.K. (1999)

Bruner II, G. C., Hensel, P. J., James, K. E.: Marketing scales handbook: a compilation of multi-item measures for consumer behavior & advertising research (Vol. 4). American Marketing Association, Chicago, IL (2005)

Cheung, G., Rensvold, R. B.: Assessing extreme and acquiescent response sets in cross-cultural research using structural equations modeling. Journal of Cross-Cultural Psychology, 31, 187-212 (2000). doi:10.1177/0022022100031002003.

Davidov, E., De Beuckelaer, A.: Testing the equivalence of an instrument to assess Schwartz's human values: how harmful are translations? International Journal of Public Opinion Research, **22**, 485-510 (2010) doi:10.1093/ijpor/edq030.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., Billiet, J.: Measurement equivalence in cross-national research. Annual Review of Sociology, **40**, 55-75 (2014) doi:10.1146/annurev-soc-071913-043137.

Davis, C. M. (1998): Handbook of sexuality-related measures. Sage, Thousand Oaks, CA (1998)

De Beuckelaer, A., Weijters, B., Rutten, A.: Using ad hoc measures for response styles: a cautionary note. Quality and Quantity, **44**, 761-775 (2010) doi: 10.1007/ s11135-009-9225-z.

De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J. P., Baumgartner, H.: Using item response theory to measure extreme response style in marketing research: global investigation. Journal of Marketing Research, **45**, 104-115 (2008) doi:10.1509/jmkr.45.1.104.

Dolnicar, S., Grün, B.: Cross-cultural differences in survey response patterns. International Marketing Review, **24**, 127-143 (2007) http://dx.doi.org/10.1108/02651330710741785.

Eid, M., Lischetzke, T., Nussbeck, F. W.: Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), Handbook of multimethod measurement in psychology (pp. 283-299). American Psychological Association, Washington, DC (2006) http://dx.doi.org/10.1037/11383-020.

Harzing, A.-W.: Response styles in cross-national survey research: a 26-country study. International Journal of Cross Cultural Management, **6**, 243-266 (2006) doi:10.1177/1470595806066332.

Hirschfeld, L. A., Gelman, S. A.: Mapping the mind: domain specificity in cognition and culture. Cambridge University Press, New York, NY (1994)

Hui, C. H., Triandis, H. C.: Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, 20, 296-309 (1989) doi:10.1177/0022022189203004.